AD_____

Award Number: W81XWH-08-2-0098

TITLE: Breast Cancer Diagnostics Based on Spatial Genome Organization

PRINCIPAL INVESTIGATOR: Thomas Misteli

CONTRACTING ORGANIZATION: Geneva Foundation
Tacoma, WA 98402

REPORT DATE: July 2012

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
July 2012	Final	1 September 2008 – 30 June 2012

4. TITLE AND SUBTITLE

Breast Cancer Diagnostics Based on Spatial Genome Organization

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-08-2-0098

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Thomas Misteli
Karen Meaburn

E-Mail: mistelit@mail.nih.gov

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Geneva Foundation
Tacoma, WA 98402

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Genomes are non-randomly organized within the cell nucleus. Importantly, individual gene loci undergo changes to their spatial position during disease, including cancer. We exploited, for the first time, these changes in spatial gene positioning as a novel tool for the detection of invasive breast cancer. Repositioning events in cancer cells are not a reflection of a global spatial genome reorganization and are instead gene-specific. Out of 20 genes we examined, 8 occupied significantly different intra-nuclear positions in breast cancer compared to normal tissues. Such genes serve as Gene Positioning Biomarkers (GPBs). We have demonstrated that these GPBs can accurately distinguish normal from cancer tissues. Crucially for a diagnostic test, these repositioning events were specific to cancer. There is little variation in the spatial positioning patterns of genes amongst normal individuals and most genes did not reposition in the benign breast diseases hyperplasia and fibroadenoma. We also adapted the approach to the requirements in a clinical setting by developing a normalized standard reference distribution for all GPB genes and by demonstrating that multiplexed combinatorial gene markers can further improve sensitivity and specificity. Our observations provide the first proof-of-principle that the spatial positioning of the genome can be used for diagnostic applications.

15. SUBJECT TERMS
Breast Cancer; Cancer Diagnosis; Spatial Gene Positioning; Nuclear Architecture.

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
U	U	U	UU	19	19b. TELEPHONE NUMBER *(include area code)*

Table of Contents

Introduction

Breast cancer is one of the most prevalent cancers in the United States, with over 230,000 newly diagnosed cases and ~40,000 deaths per year [1]. Definitive diagnosis of breast cancer following indicative symptoms, such as a palpable lump or an abnormality detected by mammography, relies on histological analysis of biopsies, which uses subjective morphological and histological criteria. This subjectivity makes reliable diagnosis difficult. Definitive and accurate diagnosis is limited by the scarcity of general molecular breast cancer markers. During the DoD Breast Cancer Idea Award funding period, our objective was to test the feasibility of using differential non-random spatial organization of the genome between normal and cancer tissue as the basis of a novel, molecularly defined, diagnostic test. This method is based on the recent realization that genomes are non-randomly organized within the three-dimensional space of the cell nucleus [2, 3]. Entire chromosomes and individual genes occupy preferential positions within the interphase nucleus [2, 3] and these positions can be altered in cancer [4, 5]. For example, in pancreatic cancer, chromosome 8 shifts to a more peripheral location [6]. Similarly, chromosomes 18 and 19 change nuclear location in several cancers types, including cervical and colon [7]. In pilot studies to the DoD Breast Cancer Idea Award, we identified 4 out of 11 tested genes (*AKT1*, *VEGF*, *BCL2* and endogenous *ERBB2*) that significantly changed their position during carcinogenic transformation in an established mammary epithelial cell *in vitro* model of early breast cancer [10], where over-expression of ErbB2 in 3D cell cultures of MCF-10A cells induces a phenotype that closely mimics *in vivo* early breast carcinogenesis [8, 9]. Using our DoD Breast Cancer Idea Award, we have extended this study to human breast tissues, identifying genes that are differentially positioning in breast cancer, and, for the first time, exploit these changes in the spatial organization of the genome as an indicator of cancerous transformation, to form the bases of a novel breast cancer diagnosis strategy.

Body

The aim of this project is to identify genes which occupy distinct intra-nuclear positions in normal and malignant cells and to explore the usability of these markers for diagnostic purposes. To this end, we optimized fluorescent *in situ* hybridization (FISH) methods to detect individual genes in 4-5μm thick formalin fixed, paraffin embedded human breast tissue sections. The radial position of a gene, normalized to the size of the nucleus, was determined using a modified version of a previously developed image analysis method [10, 11, (ref 11 is included as an appendix for a more detailed account of methods and results)]. Modifications were made to the original software to account for the fact that nuclei in tissues and cancer are not always of regular elliptical shape. To account for this, the binary Euclidean distance transform (EDT) was computed for each nucleus. The EDT is a morphological operation that assigns each pixel in a nucleus a value that equals the shortest distance to the edge of the nucleus. To account for variations in nuclear size, the EDT of the geometric gravity center of a FISH signal is normalized to the maximum EDT value for the given nucleus. Using this method, no assumption regarding nuclear shape is made when determining the radial position of a gene, allowing accurate comparisons between tissues, even if there are differences or irregularities in nuclear shape or size. All alleles in a nucleus are included in the analysis and nuclei are included regardless of the number of alleles present, unless no gene signals were present in a nucleus. For each sample, data from 88-220 nuclei, which were acquired from multiple randomly selected regions of the

tissue sample, were analyzed and combined to determine the cumulative relative radial distribution (RRD) for each gene in a tissue. The RRD is a standard measure of a genes position in a population and is defined as the statistical distribution of the radial position of all alleles in a cell population. The distribution of a gene's position within a tissue can either be expressed as a distribution graph or as a cumulative frequency graph (Fig. 1). RRDs were statistically compared to each other using the two-sample 1-D Kolmogorov-Smirnov test (KS; $P < 0.01$). We analyzed the RRDs of 20 genes in a panel of breast tissues made up of invasive breast carcinomas, benign diseased tissues (fibroadenoma and hyperplasia) and normal breast tissues [11]. To enable an unbiased screening approach, the 20 genes were selected randomly and irrespective of their function, and mapped to a range of different (14) chromosomes.

Identification of putative cancer markers

We individually cross-compared the RRD of a gene in each cancer tissue, to the RRD of the given gene in each normal tissue (Table 1, [11]). Using a panel of 14 cancer tissues and 11 normal tissues, we identified 8 genes that occupied significantly different intra-nuclear positions in breast cancer compared to normal tissues (*HES5*, *HSP90AA1*, *TGFB3*, *MYC*, *ERBB2*, *FOSL2*, *CSF1R* and *AKT1*). These genes represent putative positioning-based markers of breast cancer [11] and we refer to these genes as gene positioning biomarkers (GPBs). One gene in particular, *HES5,* was highly promising as it repositioned in 91% of the pair-wise comparisons (83/91). The observed repositioning events were not the consequence of non-specific global spatial genome reorganization within cancer nuclei as indicated by the fact that only a minority of the tested genes underwent significant repositioning (8/20 genes; Table 1, [11]). Genomic instability is prevalent in cancer, and it is possible that the altered copy number of a given gene could influence positioning patterns. However, repositioning did not correlate with changes to the copy number of a given gene, nor with the degree of genomic instability we detected within a given tumor [11]. Thus, we conclude that the apparent cancer specific repositioning events were also not due to genomic instability [11].

We initially thought that the 3D culture model of breast cancer may be useful to screen genes for maker potential before taking promising markers onto tissues. However, of the 4 genes that repositioned in this system (*AKT1*, *VEGF*, *BCL2*, *ERBB2*) [10], only 2 repositioned in breast cancer tissues (*AKT1*, *ERBB2*) (Table I and II) [11]. Moreover, *TGFB3* which did not reposition in the cell culture model, repositioned in the majority of breast cancer tissues (Table I and II) [10, 11]. Thus, we directly tested new candidate genes for their cancer repositioning potential on breast tissues.

Sensitivity

Our identification of potential marker genes relied on the comparison of gene position in a set of cancer tissues compared to a set of normal tissues. Although useful for the *bona fide* identification of putative markers, the requirements in a clinical diagnostic setting are very different. In a clinical setting, an unknown sample must be classified as normal or cancerous often in the absence of control tissues from the patient. For a useful diagnostic test, it must be unambiguous if the gene has repositioned or not, which is not always possible when comparing to multiple normals. To move our markers closer to the more realistic clinical setting we developed a standardized normal distribution (SND) for each gene of interest. To establish a SND we pooled positioning data for individual genes from 6-8 normal tissues [11]. We

compared the position of genes in our known cancer samples to the gene's SND, and the position of these genes was able to accurately distinguish cancer tissues from normal in 82% (*HSP90AA1*) to 64.3% (*AKT1*) of cases, depending on the gene used (Table II, [11 and unpublished data]). Correspondingly, the false negative rate, defined as the percentage of cancer tissues exhibiting a gene distribution indistinguishable from the SND ranged from 18% (*HSP90AA1*) to 35.7% (*AKT1*). Multiplexing genes improved the sensitivity (Table III, [11]). For example, the positioning pattern of *HES5* combined with any of other of the 7 marker genes resulted in >94% of cancers being correctly identified as cancer, and any 2 markers used together resulted in ≥79% of all cancers correctly being identified as cancer.

During the course of this study, we established that the analysis of ~100 nuclei was a reliable sample size for robustly determining a genes radial position in a tissue (Fig S1 in [11]). The small sample size required is a benefit for a diagnostic setting, as it reduces the need for additional invasive procedures, and can use the remains of the biopsy sample, not needed for conventional diagnosis. The sample size for RRD was based on the analysis of normal tissues, however, cancer tissues maybe more sensitive to this because most diagnostic cancer specimens will contain a varying mixture of normal and cancer cells and we image the tissues as randomly as possible, to try to incorporate any heterogeneity within a given tumor (although we try to avoid normal tissue and connective tissue within the tumor tissue section). The normal cells would dilute any repositioning events detected for a cancer tissue, and could lead to false negatives. If this was the case, a greater number of nuclei would be required to obtain robust and truly representative RRD for a given cancer. To determine the minimal fraction of cancer cells required in a sample we performed a blinded mixing experiment. To this end, we generated datasets of 160 nuclei containing varying proportions (10%, 30%, 40%, 50%, 60%, 70% and 100%) of cancer nuclei [11]. As a source of nuclei for these mixed datasets, we first created master datasets of 200 normal and 200 cancer nuclei, which contained *HES5* signals and had been randomly selected from multiple tissues, and used these master datasets to generate the mixed ratio datasets. The RRD of *HES5* was determined for each of the mixed ration datasets using our standard procedure, and compared to our SND. Differential positioning of *HES5* could be detected in datasets containing up to 40% normal nuclei ($P \leq 0.001$) (Fig S1 in [11]), demonstrating that tissue heterogeneity does not preclude accurate detection of gene position and identification of cancer tissues [11].

Specificity

For a marker to be of clinical use it must have a low false positive rate, to reduce misdiagnosis and subsequent unnecessary treatment and burden on individuals who do not have cancer. Since the genome can reorganizes in diseases other than cancer [12, 13], and some genomic loci are differently positioned in proliferating compared to non-proliferating cells [10, 13, 14], it is possible that some of the repositioning events identified are not specific to cancer, but would also be detected in benign disease. Another explanation for differences in gene positioning might be variability of the location of a gene amongst individuals. To rule out these trivial explanations of the repositioning, and to eliminate the GPBs that have a high false positive rate, the radial positioning patterns of our top 8 marker genes were compared between normal tissues, and positioning patterns were also compared between benign disease (fibroadenoma and hyperplasia; not including atypical hyperplasia, which is linked to breast cancer development) and normal tissue (Table II, [11]). There was a low-degree of variability in spatial gene positioning between

individuals, ruling out that the observed repositioning events in tumor samples are due to random variations in positioning patterns amongst individuals (Table II, [11]). Moreover, most genes did not reposition in benign disease (Table II, [11]). The notable exception was *ERBB2*, which repositioned in 3/5 (60%) of benign tissues (Table II, [11]). The false positive rate, defined as the percentage of non-cancer (normal and benign) tissues exhibiting a gene distribution significantly different to the SND, ranged from 0% (*FOSL2*, *TGFB3*, *MYC* and *CSF1R*) to 18.2% (*HSP90AA1*), depending on the gene (Table II, [11]). Again the exception was *ERBB2*, which had a false positive rate of 28.6%, thus, we have eliminated the radial position of *ERBB2* as a promising diagnostic marker.

The sensitivity and specificity of the GPBs are similar or below the error rates in commonly used morphologically based diagnostic methods, although obviously need to be validated in larger numbers of tissues. Taken together, we have identified several GPBs, which are able to distinguish cancerous tissue from normal and benign diseased tissue with high accuracy. Our observations provide the first proof-of-principle that the spatial positioning of the genome can be used for diagnostic applications [10, 11].

Validation of markers in larger sample sets

We next focused on two major areas: 1) validation of markers in larger samples sets and 2) development of high-throughput imaging and analysis methods to enable the analyses of large sample sets, including allowing the comparison of positioning patterns between various breast tumor types. Moreover, these analysis methods are required in the clinical setting, if this type of diagnostic test is to be a practical diagnostic method.

Our initial studies identified gene positioning markers based on analysis of ≤14 cancer samples. While we obtained statistically significant results and were able to identify candidate markers, the robustness of the markers needs to be tested on larger datasets, ideally containing hundreds or thousands of samples. To address the issue of tissue numbers we have initiated the use of tissue microarrays (TMAs). A TMA is an array of small cores of tissues (typically 0.6-2mm in diameter) placed on a single glass microscopy slide. A typical array contains between 50-150 individual samples. The advantage of this approach is that several hundred samples can be simultaneously processed for FISH and imaging. The approach required optimization of FISH and imaging conditions. We have now implemented standardized conditions for FISH on TMAs (Fig. 2a). We can now routinely stain and image TMAs from various sources (US Biomax, NCI, Aureon Pharmaceuticals). Moreover we have established that typical cores on a TMA contain a sufficient number cells for gene positioning analysis. An additional benefit of many TMAs is that multiple cores from the same individual are present on the slide. Utilizing this, we addressed the issue of possible heterogeneity within tumors by use 2 or more different cores of tissue from the same tumors. Importantly for a diagnostic test, we found similar positioning patterns of genes between the different cores of the same individual (Fig 2b). A limitation of TMAs is that not all tissue cores on a TMA are present or usable (due to damage or because they mainly contain stroma) on every slide but having multiple tissues of a given individual increases the number of tissues which can be analyzed per TMA.

A major bottleneck in the analysis of these samples is the image analysis. While we have relied previously on a semi-manual method to identify nuclei and FISH signals in the tissue sample, the

large number of individual nuclei (100 per tissue sample, several hundred tissues) exceeded our analysis capacity. In this semi-automated analysis nuclei must be manually identified and segmented. To overcome this problem we have, in close collaboration with Dr. Stephen Lockett, NCI, developed and implemented a novel image analysis tool. In this approach, nuclei are detected automatically using a new imaging software tool. Detection of FISH signals is done using an already established imaging tool, called NMFA-FLO (Nuclei Manual and FISH automatic). In order to achieve accurate segmentation of nuclei in tissue we used an artificial neuronal network (ANN)-based supervised pattern recognition approach to screen out well segmented nuclei, after image pre-processing and multistage watershed segmentation (Fig. 3) [15, 16; Refs 15and 16 are included in the appendix to give a fuller details]. In this approach we provide the software with a training set of images of manually identified nuclei. The software analyzes the features of these objects and develops an internal algorithm to identify well segmented nuclei based on a combination of 64 mathematical features. This fully automated approach identifies nuclei with ~ 80% accuracy and, importantly, with very low false positive rates. In addition to the ANN pattern recognition selection of well segmented nuclei, we have recently implemented the use of a ranked retrieval for nuclei, which used uses logistic regression to output the probability of a nucleus being correctly segmented [17]. This ensures only the best, and very accurately, segmented nuclei are used for gene positioning analysis. Accurate analysis of gene positioning is highly dependent on accurate segmentation of nuclei.

The task of fully automating nuclear segmentation was more difficult that initially anticipated, and many approaches had to be tested to develop a usable and robust automatic nuclear segmentation tool. Nuclear segmentation in tissues is difficult because nuclei tend to touch each other meaning that simply nuclear boundary detection can not be done by using the difference in signal intensity between the background and nuclei. There is also considerable variation in morphology and “texture” (variation in DAPI intensities throughout the nuclei) between normal and cancer tissues (e.g. Fig. 1), and between individual cancer tissues. Thousands of nuclei segmented by the software have been manually checked, to ensure the software is correctly calling well segmented nuclei, and to help improve the segmentation and to help teach the pattern recognition software to more accurate identify well segmented nuclei. Using this approach we compared the accuracy of our diagnostic method using the newly developed fully automated nuclei/FISH detection system and our previously used manual method. We find comparable results (Fig. 4). We are still in the process of optimizing these algorithms, particularly to work well on normal tissues, but are close to having a fully automated image analysis tool in hand. We have also been running ~100 breast cancer tissues through the automated analysis software (with the genes *HES5* and *FOSL2* visualized by FISH), which were not used during the development of the software, to ensure the software’s robustness. So far, comparisons to manual analysis result in ~80% agreement on if a gene has repositioned or not in a cancer tissue, compared to the SND (21/26 tissues) (Fig 4). This tool now puts us in a position to analyze large datasets. The analysis of large datasets will allow us to not only validate our GPBs for diagnostic purposes, but allow us to start addressing questions of a more prognostic nature, such as whether markers are tumor-type specific or correlate with outcome or prognosis.

Table I. Identification of cancer markers

Gene	Number of cross-comparisons between individual normal and cancer tissues			Number of cross-comparisons between individual normal tissues		
	SD	**Total**	**% SD**	**SD**	**Total**	**% SD**
HES5	83	91	91.2	1	21	4.8
MYC	47	66	71.2	0	15	0.0
FOSL2	58	91	63.7	1	21	4.8
HSP90AA1	41	66	62.1	0	15	0.0
CSF1R	56	91	61.5	1	21	0.0
ERBB2	73	126	57.9	7	36	19.4
AKT1	54	98	55.1	5	21	23.8
TGFB3	59	112	52.7	0	28	0.0
HES1	6	12	50.0	1	3	33.3
ZNF217	3	6	50.0	1	1	0.0
VEGFA	10	24	41.7	0	6	0.0
MMP1/3/12	2	6	33.3	0	1	0.0
CCND1	7	24	29.2	0	6	0
PTGS2	1	4	25.0	n.d.	n.d	n.d
BCL2	4	18	22.2	1	3	33.3
HEY1	1	15	6.7	0	3	0.0
BRCA1	0	2	0.0	n.d.	n.d	n.d
PTEN	0	2	0.0	n.d.	n.d	n.d
TLE1	0	2	0.0	n.d.	n.d	n.d
TJP1	0	1	0.0	n.d.	n.d	n.d
Total	**505**	**857**	**58.9**	**18**	**201**	**9.0**

SD = significantly different; based on 1-D KS-test, $P < 0.01$. n.d. = not determined. This table is taken directly from [11].

Table II. Single gene true positive, false positive and negative rates

Gene	True positive	False negatives	Normal tissue false positives	Benign disease tissue false positives	Total false positives
HES5	29/40 (72.5%)	11/40 (27.5%)	0/7 (0%)	1/6 (16.7%)	1/13 (7.7%)
FOSL2	29/39 (74.4%)	10/39 (74.4%)	0/7 (0%)	0/6 (0%)	0/13 (0%)
HSP90AA1	9/11 (81.8%)	2/11 (18.2%)	0/6 (0%)	2/5 (40%)	2/11 (18.2%)
TGFB3	11/14 (78.5%)	3/14 (21.4%)	0/8 (0%)	0/5 (0%)	0/13 (0%)
MYC	8/11 (72.7%)	3/11 (27.3%)	0/6 (0%)	0/5 (0%)	0/11 (0%)
ERBB2	10/14 (71.4%)	4/14 (28.6%)	1/9 (11.1%)	3/5 (60%)	4/14 (28.6%)
CSF1R	9/13 (69.2%)	4/13 (30.8%)	0/7 (0%)	0/5 (0%)	0/12 (0%)
AKT1	9/14 (64.3%)	5/14 (35.7%)	1/7 (14.3%)	0/5 (0%)	1/12 (8.3%)
Total	98/129 (76%)	25/103 (24.3%)	2/57 (3.5%)	6/40 (15.0%)	8/97 (8.2%)

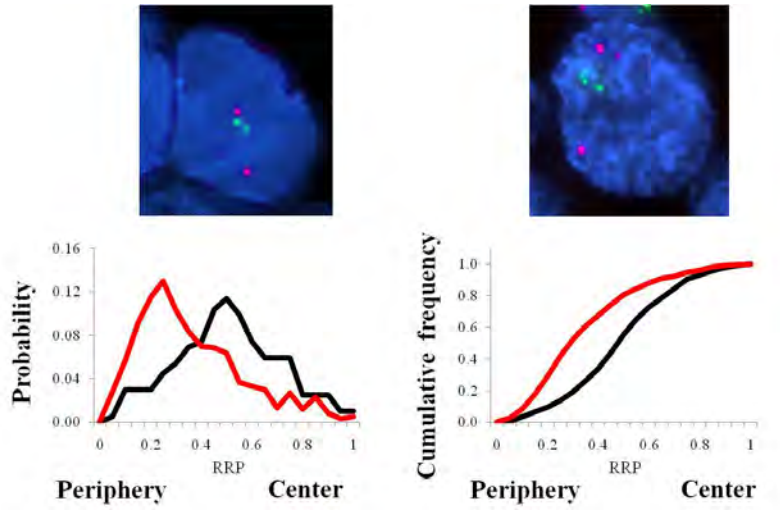
The number (and percentages) of tissues that gives either a false negative, false positive or true positive result. For a false negative, a gene has a similar RRD in a cancer tissue to that of the pooled normal distribution (1-D KS-test; $P > 0.01$). A false positive is scored when a gene has a statistically different RRD to that of the pooled normal in non-cancerous breast tissues ($P < 0.01$), and a true positive is scored when a gene has a statistically different RRD to that of the pooled normal in cancerous breast tissues. This table has been adapted from Table IV in [11] to include true positive data, and to include additional, unpublished, data.

Table III. Use of multiple markers

	HES5	*HSP90AA1*	*TGFB3*	*FOSL2*	*MYC*	*ERBB2*	*CSF1R*	*AKT1*
HES5	X	13/13 (100%)	14/14 (100%)	36/38 (94%)	13/13 (100%)	14/14 (100%)	14/14 (100%)	14/14 (100%)
HSP90AA1		X	12/13 (92%)	13/14 (93%)	11/11 (100%)	13/13 (100%)	13/13 (100%)	12/14 (86%)
TGFB3			X	13/14 (93%)	12/13 (92%)	13/14 (93%)	14/14 (100%)	13/14 (93%)
FOSL2				X	12/13 (92%)	12/14 (86%)	11/13 (85%)	11/14 (79%)
MYC					X	11/14 (79%)	11/13 (85%)	13/14 (93%)
ERBB2						X	11/14 (79%)	12/14 (86%)
CSF1R							X	12/14 (86%)
AKT1								X

The number (and percentage) of cancers where at least one of the indicated pair of genes repositioned, compared to the pooled normal distribution (1-D KS-test, $P < 0.01$). Red boxes indicate a 100% detection rate, pink = > 90%, yellow = > 80% and green = > 70% detection rate respectively. This table has been adapted from Table VI in [11] to include additional, unpublished, data.

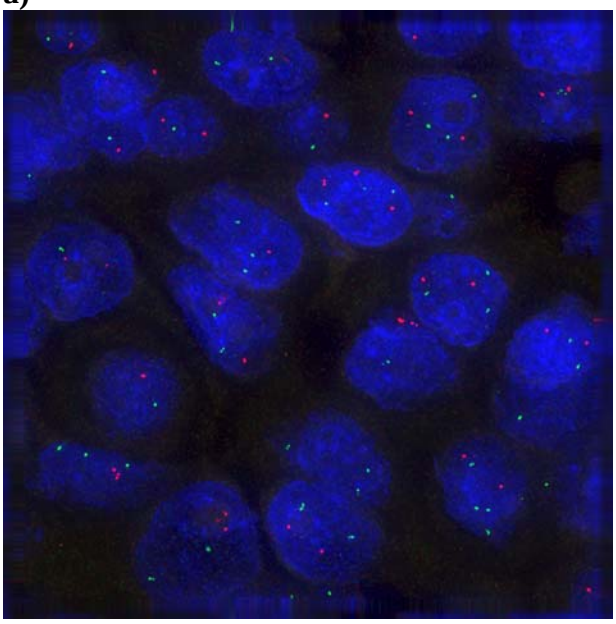
Figure 1. Differential radial positioning of a gene as a diagnostic read-out



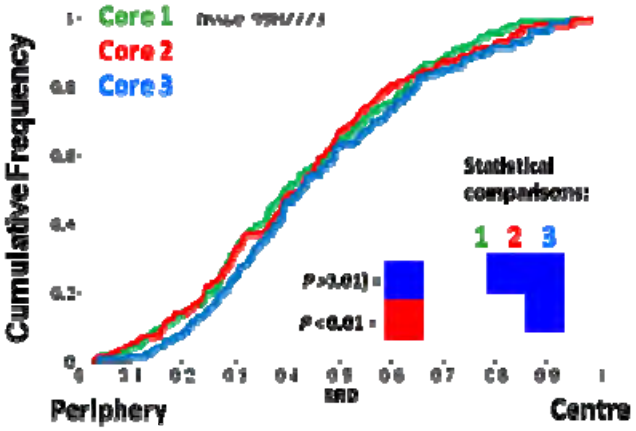
Genes that change their radial nuclear position during carcinogenesis may serve as potential diagnostic markers. (Top) FISH detection of *MYC* (red) and *ERBB2* (green) in a normal cell (left) and breast cancer (right). DNA (blue). (Bottom) The radial position of a gene can be expressed as a frequency distribution (left) or a cumulative distribution (right). Distribution of *MYC* in a normal (black) or in a breast cancer tissue (red). N =150 nuclei.

Figure 2. FISH on TMAs

a)



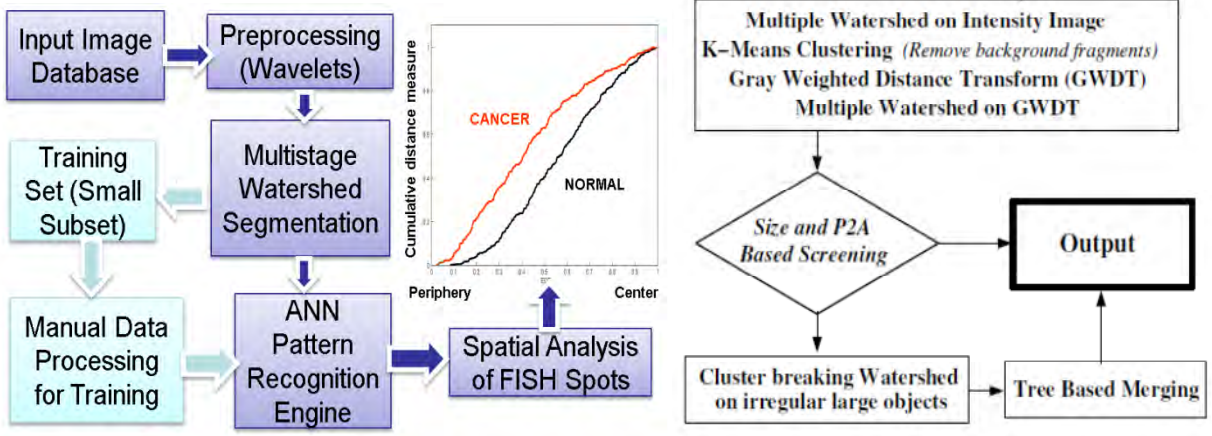
b)



a) FISH detection of *HES5* (red) and *FOSL2* (green) in nuclei (blue) on a TMA core of breast cancer. Gene signals can be detected as efficiently as in individual tissue samples. b) Gene positioning (cumulative distribution of FISH signals) is highly similar between multiple TMA tissue cores of the same tissue. Distributions of *FOSL2* gene signals are shown for 3 tissue cores from the same tumor (1-D KS-test, $P < 0.01$). A representative tissue is shown. N ≈130 nuclei.

Figure 3.

a **b**



a) The processing pipe line for the automatic analysis software. b) A scheme of the multistage watershed segmentation used to process images for automated nuclear segmentation. Part a) has been adapted from [15] and b) from [16].

Figure 4. Comparison of automatically and manually detected cancer samples

a)

Dataset	*FRA2*	*HES5*
B5/A5	0.02	0.64
B7	0.51	0.41
B12	0.09	0.02
C7	0.55	0.92
E4	0.64	0.39
E8	0.86	0.61
F2	0.94	0.13
F8	0.56	0.41
G1	0.79	0.56
G5	0.76	0.55
H5	0.80	0.29
I7	0.92	0.08
J1	0.71	0.47

b)

	Manual		HT software		Match?	
Dataset	*FRA2*	*HES5*	*FRA2*	*HES5*	*FRA2*	*HES5*
B5/A5	0.113	0.136	0.192	0.132		
B7	0.269	0.000004	0.436	0.000000		
B12	0.005	0.0003	0.00009	0.000000		
C7	0.000000	0.00004	0.000098	0.000003		
E4	0.024	0.000000	0.00005	0.000000		
E8	0.0004	0.523	0.018	0.502		
F2	0.000000	0.072	0.000001	0.001		
F8	0.0002	0.302	0.000000	0.358		
G1	0.000000	0.000001	0.000000	0.000000		
G5	0.000136	0.016	0.0005	0.020		
H5	0.000084	0.000003	0.0007	0.029		
I7	0.0009	0.00001	0.0005	0.013		
J1	0.000000	0.053	0.0004	0.983		*

The accuracy of cancer detection was compared between the semi-manual (manual) analysis method and the fully automated high throughput (HT) software. a) Gene positioning for the genes *HES5* and *FRA2* (*FOSL2*) was performed on the same set of images using the manual and fully automated analysis methods, for 13 tissue cores (12 of which were breast cancer samples, and 1 core (B5/A5) was from a benign tumor). The cumulative radial gene signal distribution was then compared between the 2 methods for the same tissue, using the 1-D KS-test (*P* values shown). In most cases the distributions between the 2 analysis methods are highly similar (blue). b) The positioning patterns generated from both analysis methods were then compared to the SND (generated by the manually analysis method), using the 1-D KS-test (*P* values shown). Red denoted significant difference of gene positioning in a tissue ($P < 0.01$) to the SND. In most cases, the classification of significantly different to the PND or not, (thus being classed as cancer or not) was the independent of the nuclear segmentation method used (green boxes). However, in 5 instances the call was different when the fully automated software was used.

Key research accomplishments

- The interphase spatial positioning patterns of 20 genes have been screened in a panel of normal and invasive cancer human breast tissues to identify candidate marker genes for breast cancer detection.

- Demonstration of little variation in the spatial position of a given gene amongst normal individuals. Thus, any repositioning between normal and cancer tissues are specific to disease and are not a consequence of inter-individual differences in positioning patterns.

- Demonstration of gene-specific repositioning events associated with carcinogenesis.

- Demonstration of an absence of global genome reorganization in cancer cells.

- Identification of 8 potential cancer maker genes (*HES5*, *MYC*, *FOSL2*, *HSP90AA1*, *CSF1R*, *ERBB2*, *AKT1* and *TGFB3*), since they reposition in the majority of analyzed tumors.

- Demonstration that the repositioning events in cancer are not a consequence of genomic instability.

- Establishment of a standard normal distribution for comparison with unknown samples.

- Validation that repositioning events are specific to cancer, and not a general disease response, with the exception of *ERBB2*, which we have ruled out as a promising cancer-specific marker.

- Determination of false positive/negative rates.

- Demonstration of suitability of multiplexed combinatorial gene markers.

- Development and implementation of high-throughput FISH methods using TMAs for analysis of large sample sets.

- Development of novel image segmentation methods based on neuronal network analysis, and including ranked retrieval for nuclei assessment, which uses logistic regression to output the probability of a nucleus being correctly segmented.

Reportable outcomes

Patent application:

2008/9 Method for detection of cancer based on spatial genome organization. E-283-2008/0--PCT-02

Publications:

2009 Meaburn KJ, Gudla, PR, Khan S, Lockett S and T Misteli. Cancer Detection Based on Spatial Genome Organization. *The Journal of Cell Biology, 187, 801-812*
This publication was widely covered in the international popular press including on *Fox News, Yahoo,* and *The Independent of London* etc.

2009 Nandy K, Gudla PR, Meaburn KJ, Misteli T, Lockett SJ. Automatic nuclei segmentation and spatial FISH analysis for cancer detection. *EMBS.2009:6718-21*

2012 Nandy K, Gudla P. R., Amundsen R., Meaburn K. J. Misteli T, Lockett S.J. Automatic segmentation and supervised learning based selection of nuclei in cancer tissue images. Cytometry Part A. *In press.*

2012 Cukierski W.J., Nandy K, Gudla P.R., Meaburn K.J., Misteli T, Foran D.J., Lockett S.J. Ranked Retrieval of Segmented Nuclei for Objective Assessment of Cancer Gene Repositioning. *Under review at BMC Bioinformatics*

Meeting Presentations:

Oral presentations:

- Center for Excellence in Chromosome Biology: Post-doctoral retreat, NIH, Bethesda, MD. USA. 17 Dec. 2010
- Ventana Symposium, Tucson, AZ. 9-13 March, 2010
- Microscopic Image Analysis with Applications in Biology. Bethesda, MD. 3-4 Sept. 2009
- 17th International Chromosome Conference, Boone, NC, USA. 23-6 June 2009
- FASEB Summer Research Conference on Nuclear Structure and Cancer, Saxton River, VT, USA. 14-19 June 2009
- Marie Curie Higher Order Genome Architecture Meeting, Edinburgh, UK. 1-5 April 2009

- Departmental seminar program. Biosciences, Brunel University, Middlesex, UK. 30 March 2009
- 11th European Workshop on Cytogenetics and Molecular Genetics of Solid Tumours. Bilbao, Spain. 6-9 Sept. 2008

Poster presentations:

- 6th Era of Hope Conference, DoD/BCRP Conference, Orlando, FL. 2-5 Aug. 2011

- 75th Symposium: Nuclear Organization & Function meeting, Cold Spring Harbor Laboratory, NY. 2-7 June 2010

- Center of Excellence in Chromosome Biology (CECB), Center for Cancer Research, National Cancer Institute symposium "Chromatin Dynamics in Development and Disease". 8-9 April 2010

- Center of Excellence in Chromosome Biology: Postdoctoral Fellows Retreat. Bethesda MD. 15 Dec 2009
- National Cancer Institute (NCI) Translational Science Meeting. Washington, DC. 7-9 November 2008

- NCI Symposium on Chromosome Biology: Genome-Wide Chromatin Structure and Function, Bethesda, MD. 30-31 Oct 2008

- NIH Research Festival, Bethesda, MD. 14-17 Oct 2008

Funding applications based on work supported by the grant:

Department of Defense, Breast Cancer Idea Expansion Award
Project: Breast cancer early diagnostics and prognostics based on interphase spatial genome positioning – Misteli (PI). 07/01/2012-06/30/2014.
The aim of this project is to expand the work of the current idea award to develop GPBs for the detection of early breast cancer, for the staging of breast cancers and as predictive indicators of recurrence, metastasis and survival.

Department of Defense, Prostate Cancer Idea Development Award.
Project: Prostate cancer diagnostics and prognostics based on interphase spatial genome positioning – Misteli (PI). 07/01/2012-06/30/2015.
The aim of this project is to test the feasibility of using spatial genome organization patterns as indicators of prostate cancer and to identify gene markers for use in diagnostic applications. This project uses the same technology and synergized with the current funded award but focuses on the identification of markers in prostate cancer.

Personnel (not salaries) receiving pay from the research effort:
None

Conclusions

We have developed a strategy to identify novel cancer biomarkers, based on the differential spatial localization of genes within the cell nucleus. Application of this strategy has led to the characterization of 8 promising novel cancer biomarkers. We have tested their usefulness in a test set of human breast formalin-fixed paraffin embedded tissues, which include cancerous, normal and benign tissues. We have adapted the approach to the requirements in a clinical setting by developing a normalized standard reference distribution for all promising marker genes. Using this approach, cancer tissue can reliably be detected with high accuracy. Moreover, we have developed methods to apply this approach to large sample sets, in a high-throughput fashion. These observations are proof-of-principle for the application of spatial genome positioning as a novel approach in cancer diagnosis and the recently developed tools provide the basis for the systematic analysis of cancer samples.

In the long-term, these efforts should lead to the development of a robust, standardized method for the detection of breast cancer in a routine diagnostic laboratory setting. Analysis of gene positioning patterns promises to be a sensitive and effective diagnostic approach for breast cancer. Gene positioning has the potential for very early detection since genome reorganizations can occur prior to physiological or pathological changes [18]. Spatial positioning patterns also have the promise to stratify subtypes of breast cancer and to act as robust prognostic markers, since gene expression patterns are influence by a loci's spatial position [2, 3]. Consistently, we find differences in gene positioning patterns between individual breast cancers [11]. Our approach overcomes several of the limitations of current and currently proposed diagnostic tests since it is i) highly quantitative, ii) based on single cell analysis, iii) applicable to extremely small tissue samples, thus reducing the requirement for additional exploratory invasive procedures, iv) is independent from the generation of metaphase chromosome, which can be difficult to obtain from solid tumors and v) is insensitive to protein and RNA degradation, which commonly occurs during biopsy sample handling, unlike immunohistochemistry-, PCR-, or microarray-based diagnostic approaches. Moreover, our method of diagnostics can easily be integrated into clinical laboratories as an extension to existing routine cytogenetic procedures using FISH to detect gene amplifications in solid tumors. Our assay will extend and complement conventional morphology-based diagnostics and it is anticipated that the combined use of standard pathological indicators and our method will be a highly accurate, quantitative and powerful diagnostic approach. Our recent success in the development of software to automatically analyze large dataset positions us ideally to fully exploit spatial genome organization as a novel strategy in breast cancer diagnosis, and potentially in breast cancer prognosis.

References

1. Siegel, R., et al., *Cancer statistics, 2011: The impact of eliminating socioeconomic and racial disparities on premature cancer deaths*. CA Cancer J Clin, 2011. **61**(4): p. 212-36.
2. Misteli, T., *Beyond the sequence: Cellular organization of genome function.* Cell, 2007. **128**(4): p. 787-800.
3. Ferrai, C., et al., *Gene positioning.* Cold Spring Harb Perspect Biol, 2010. **2**(6): p. a000588.
4. Zink, D., et al., *Nuclear structure in cancer cells.* Nat Rev Cancer, 2004. **4**(9): p. 677-87.
5. Meaburn, K.J. and T. Misteli, *Cell biology: chromosome territories.* Nature, 2007. **445**(7126): p. 379-781.
6. Wiech, T., et al., *Human archival tissues provide a valuable source for the analysis of spatial genome organization.* Histochem Cell Biol, 2005. **123**(3): p. 229-38.
7. Cremer, M., et al., *Inheritance of gene density-related higher order chromatin arrangements in normal and tumor cell nuclei.* J Cell Biol, 2003. **162**(5): p. 809-20.
8. Debnath, J. and J.S. Brugge, *Modelling glandular epithelial cancers in three-dimensional cultures*. Nat Rev Cancer, 2005. **5**(9): p. 675-88.
9. Muthuswamy, S.K., et al., *ErbB2, but not ErbB1, reinitiates proliferation and induces luminal repopulation in epithelial acini*. Nat Cell Biol, 2001. **3**(9): p. 785-92.
10. Meaburn, K.J. and T. Misteli, *Locus-specific and activity-independent gene repositioning during early tumorigenesis.* J Cell Biol, 2008. **180**(1): p. 39-50.
11. Meaburn, K.J., et al., *Disease-specific gene repositioning in breast cancer.* J Cell Biol, 2009. **187**(6): p. 801-12.
12. Borden, J., and L. Manuelidis, *Movement of the X chromosome in epilepsy*. Science, 1988. **242**(4886): p. 1687-91.
13. Meaburn, K.J., et al., *Primary laminopathy fibroblasts display altered genome organization and apoptosis*. Aging Cell, 2007. **6**(2): p. 139-53.
14. Bridger, J.M., et al., *Re-modelling of nuclear architecture in quiescent and senescent human fibroblasts.* Curr Biol, 2000. **10**(3): p. 149-52.
15. Nandy, K, Gulda, P. R., Meaburn, K. J., Misteli, T. and S.J. Lockett. 2009. Automatic nuclei segmentation and spatial FISH analysis for cancer detection. EMBS., 2009, 6718-21.
16. Nandy, K, Gudla, P. R., Amundsen, R., Meaburn, K. J. Misteli, T, and S. J. Lockett. *Automatic segmentation and supervised learning based selection of nuclei in cancer tissue images*. Cytometry Part A. 2012. *In press.*
17. Cukierski, W.J., Nandy, K, Gudla, P.R., Meaburn ,K.J., Misteli, T, Foran, D.J., and S.J. Lockett. Ranked Retrieval of Segmented Nuclei for Objective Assessment of Cancer Gene Repositioning. *Under review at BMC Bioinformatics*
18. Ragoczy, T., et al., *A genetic analysis of chromosome territory looping: diverse roles for distal regulatory elements*. Chromosome Res, 2003. **11**(5): p. 513-25.

Appendices

Meaburn, K. J., Gulda, P. R., Khan, S., Lockett, S.J. and T. Misteli. 2009. Disease specific gene repositioning in breast cancer. J. Cell Biol., 187, 801-812.

Nandy, K, Gulda, P. R., Meaburn, K. J., Misteli, T. and S.J. Lockett. 2009. Automatic nuclei segmentation and spatial FISH analysis for cancer detection. EMBS., 2009, 6718-21.

Nandy, K, Gudla, P. R., Amundsen, R., Meaburn, K. J. Misteli, T, and S. J. Lockett. 2012. Automatic segmentation and supervised learning based selection of nuclei in cancer tissue images. Cytometry Part A. *In press.*